# Definición del problema/objetivo de investigación

## Problemática:

La empresa "IquiqueMiami LTDA." ha observado fluctuaciones significativas en sus ingresos durante el último año fiscal. A pesar de tener una amplia variedad de productos y presencia en varias estados del país de Estados Unidos, la empresa ha enfrentado desafíos debido a la alta competitividad en el mercado y cambios en las preferencias de los consumidores. La dirección ha identificado que una estrategia de precios ineficaz podría ser una de las causas principales de estos resultados inconsistentes.

## Objetivos:

# Introducción

En los últimos años, el movimiento económico de empresas pequeñas enfrenta desafíos constantes debido a la competencia, las cambiantes preferencias de los consumidores o clientes, y la rápida evaluación de la tecnología. “Iquique Miami LTDA.”, una empresa en el sector minorista, se ha comprometido a mejorar sus ventas y la satisfacción del cliente mediante la implementación de un enfoque basado en datos.

## Descripción del conjunto de datos

### Diccionario de datos

|  |  |
| --- | --- |
| Nombre de la variable | Descripción |
| Area code | Código del area de la tienda |
| Profit | Ganancia en dolares ($) |
| Margin | Ganancia + Total Gasto ($) o Ventas - COGS ($) |
| Sales | Valores adquiridos en Ventas ($) |
| State | Estado de la tienda |
| Market | Región de la tienda |
| Market Size | Tamaño de la tienda |
| COGS | Costos de los bienes vendidos |
| Total Expenses | Gastos totales para adquirir que el producto se venda ($) |
| Marketing | Gastos en Marketing |
| Inventory | Valor del inventario del producto en el momento de la venta ($) |
| Budget Profit | Ganancia esperado ($) |
| Budget COGS | COGS esperado ($) |
| Budget Margin | Ganacia esperado + Gastos totales esperados ($) o Ventas esperadas – COGS esperado ($) |
| Budget Sales: | Valor esperado obtenido en ventas |
| ProductID | ID del producto |
| Date | Fecha de venta |
| Product Type | Tipo/Categoría del producto |
| Product | Descripción del producto |
| Type | Tipo |

### Datos cuantitativa

|  |  |
| --- | --- |
| Nombre de la variable | Tipo de variable |
| Area code | Entero |
| Profit | Entero |
| Margin | Entero |
| Sales | Entero |
| COGS | Entero |
| Total Expenses | Entero |
| Marketing | Entero |
| Inventory | Entero |
| Budget Profit | Entero |
| Budget COGS | Entero |
| Budget Margin | Entero |
| Budget Sales | Entero |
| ProductId | Entero |

### Datos cualitativos

|  |  |
| --- | --- |
| Nombre de la variable | Tipo de variable |
| State | Texto |
| Market | Texto |
| Market Size | Texto |
| Product Type | Texto |
| Product | Texto |

# Preprocesamiento de Datos

En un dataset es importante conocer las variables o tuplas que la componen, ya que esto va de la mano con la problemática de la empresa, es por esto que se debe aplicar los siguientes código de programación usando R que a su vez usa la librería Tidyverse.

En primer lugar debemos capturar o leer los datos del dataset y guardarlos en un objeto, de tal manera que permita acceder a los atributos del mismo, se debe usar el siguiente código:

|  |
| --- |
| datos <- read\_csv("sales\_oficial.csv") |

En segundo lugar, al almacenar los datos del dataset en el objeto “datos”, podemos usar funciones preestablecidas para conocer los datos del mismo, tal como se ve en la imagen X:

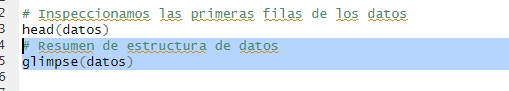


Imagen X: Inspeccionando el dataset

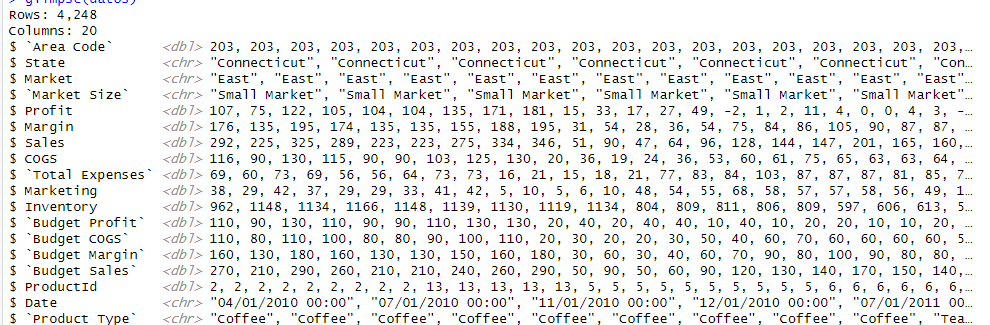
En la imagen X, se visualiza la implementación de dos códigos en R, por lo que permite conocer en breve algunos datos y los tipos de datos del dataset (por cada tupla), tal como se ve en la siguiente imagen X:  
  


Imagen X: visualizando datos

# Análisis Exploratorio de Datos

En esta sección se da a conocer en detalle la estadística de los registros existentes en el dataset, por lo que, a continuación se visualiza dos variables cuantitativas y cualitativas, lo que se pudo obtener la información a través del siguiente código R:

|  |
| --- |
| summary(datos) |

El código anterior, me muestra por cada variable su análisis estadístico, en esta oportunidad en la tabla X se muestra dos variables cuantitativas y dos cualitativas

Tabla X: descripción estadística para variables cuantitativas

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Descripción Estadística | | |
| Profit | Min | -638 | |
| 1st. Quartil | 17 | |
| Median | 40 | |
| Mean | 61.1 | |
| 3rd Quartil | 92 | |
| Max | 778 | |
| Sales | Min | | 17 |
| 1st. Quartil | | 100 |
| Median | | 138 |
| Mean | | 193 |
| 3rd Quartil | | 230 |
| Max | | 912 |

En la tabla X, muestra información relevante sobre la cantidad de registros asociadas a estas variables:

* 1. Profit:
* Min y Max: varía desde -638 hasta 778, indicando una amplia dispersión entre pérdidas y ganancias altas.
* Centralidad: La mediana de 40 indica que la mitad de las transacciones generan ganancias de 40 o menos.
* Media: Una media de 61.1, superior a la mediana, refleja un sesgo positivo debido a valores atípicos altos que elevan la media.
* Valores atípicos: Los valores atípicos, especialmente el máximo de 778, distorsionan la media hacia valores más altos.
  1. Sales:
* Min y Max: Las ventas varían desde 17 hasta 912, mostrando una gran diferencia entre las ventas más bajas y las más altas.
* Centralidad: La mediana es 138, lo cual indica que la mitad de las ventas son menores o iguales a 138. Esto sugiere que la mayoría de las ventas no son excepcionalmente altas.
* Media: La media es 193, más alta que la mediana, lo que indica un sesgo hacia valores más altos debido a ventas atípicamente grandes.
* Valores Atípicos: El valor máximo de 912 destaca como un valor atípico significativo que eleva la media por encima de la mediana.

# Modelado de Datos

# Interpretación de Resultados

* 1. Histograma de la variable *Profit*

En el siguiente gráfico consta del eje X (Profit) que representa los valores de las ganancias en dólares ($) y el eje Y (Count) representa la frecuencia o el conteo de las ventas que tienen un valor especifico de Profit, por ende la altura de cada barra indica cuantas veces ocurre un rango particular (sesgo) de ganancias en el conjunto de datos, tal como se puede ver en la imagen X:

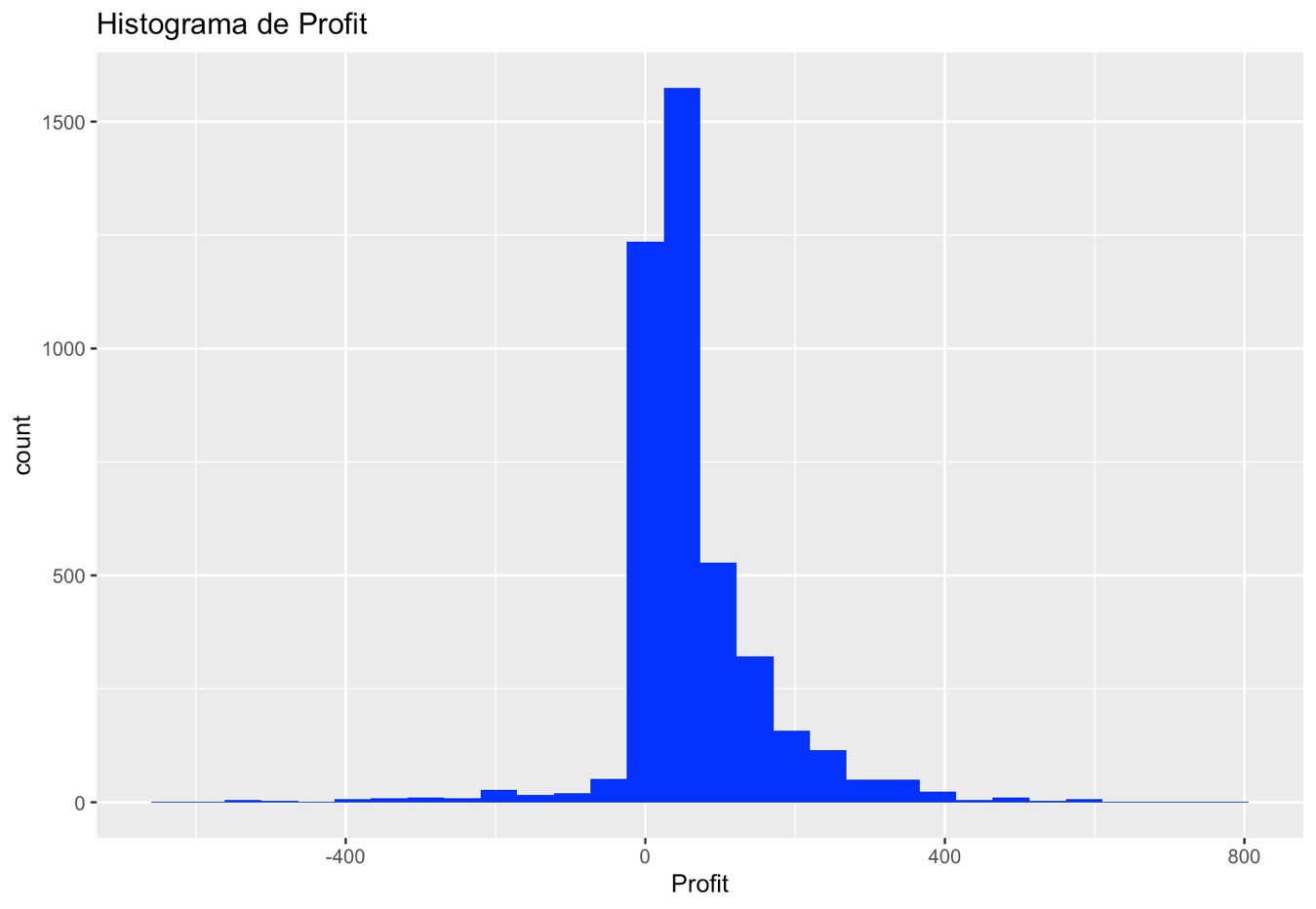


Imagen X: histograma de la variable profit

En el gráfico anterior, la mayoría de las ventas tienen ganancias bajas o cercano a cero, este sesgo de las ventas puede indicar que la empresa experimenta tanto perdidas como ganancias, sin embargo, esta distribución sesgada muestra transacciones rentables para el tipo de producto vendido.

* 1. Diagrama de cajas de *Sales*

El gráfico de cajas permite ver la dispersión de datos y validar la información descriptiva estadística de una variable numérica (Sales), esto se visualiza en la siguiente Imagen X:

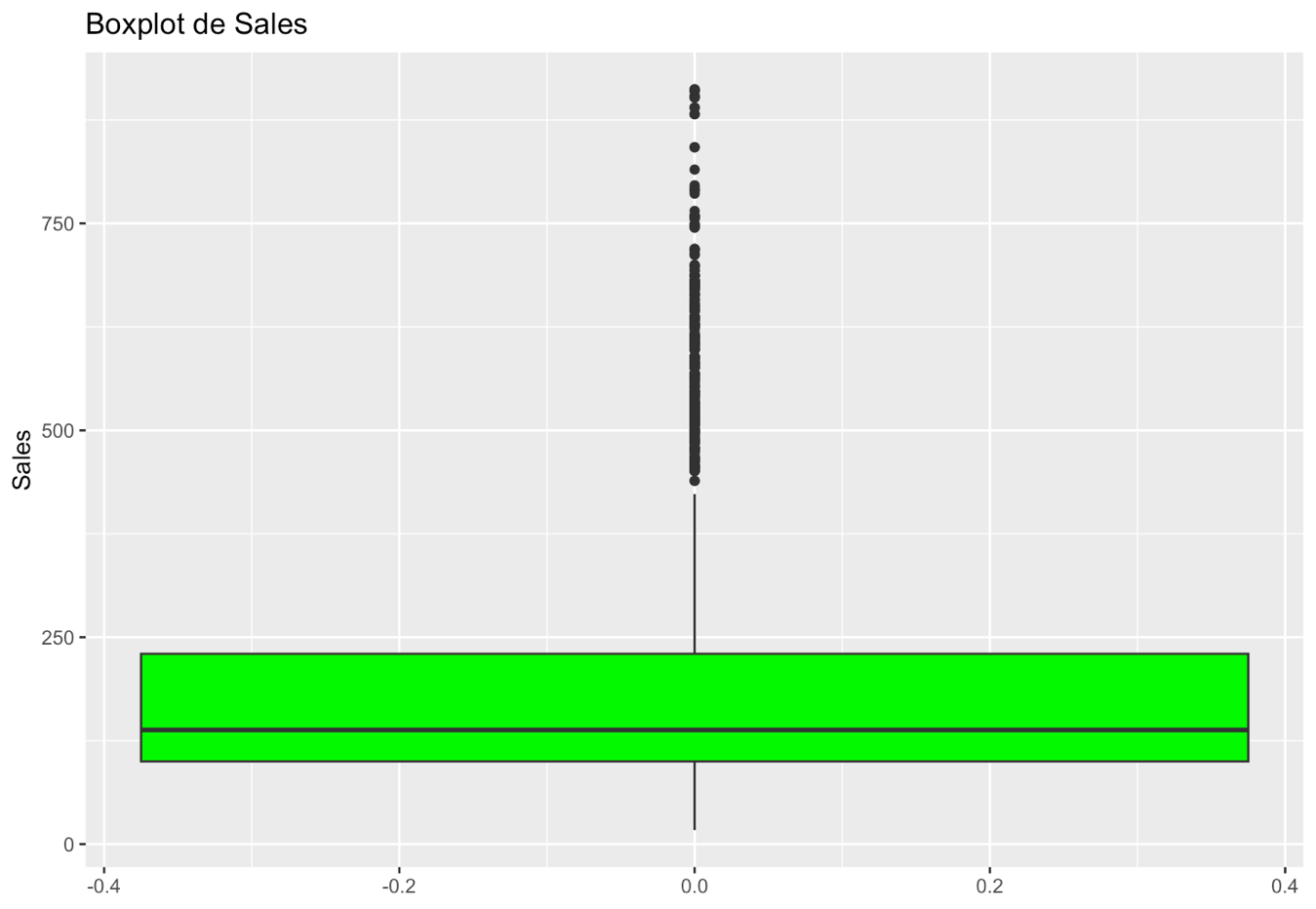


Imagen X: diagrama de cajas de la variable Sales.

En la imagen anterior se visualiza la distribución donde la mayoría de las ventas están concentradas en un rango de 100 a 230 dólares, con una mediana alrededor de 140 dólares, además existe una pronunciada variabilidad en las ventas como lo indica el amplio IQR (rango intercualictico).

# Anexos

* 1. Código R

# importación de librerias

library(tidyverse)

library(ggplot2)

#0 Definición del problema/objetivo de investigación (ficticia).

# 0.1 La empresa "IquiqueMiami LTDA." ha observado fluctuaciones significativas en sus ingresos durante

# el último año fiscal. A pesar de tener una amplia variedad de productos y presencia en varias estados

# del país de Estados Unidos, la empresa ha enfrentado desafíos debido a la alta competitividad en el

# mercado y cambios en las preferencias de los consumidores. La dirección ha identificado que una

# estrategia de precios ineficaz podría ser una de las causas principales de estos resultados

# inconsistentes.

# 1. Introducción: Descripción del conjunto de datos seleccionado y los objetivos del análisis (ficticia).

# En los últimos años, el movimiento económico de empresas pequeñas enfrenta desafíos constantes

# debido a la competencia, las cambiantes preferencias de los consumidores o clientes, y la rápida

# evaluación de la tecnología. “Iquique Miami LTDA.”, una empresa en el sector minorista, se ha

# comprometido a mejorar sus ventas y la satisfacción del cliente mediante la implementación de un

# enfoque basado en datos.

# 1.1. Diccionario de datos

# Area Code: Store's Code;

# State: Store's State;

# Market: Store's Region;

# Market Size: Store's Size;

# Profit: Profits in Dollars ($);

# Margin: Profit + Total Expenses ($) OR Sales - COGS ($);

# Sales: Values Acquired in Sales ($);

# COGS: Cost of Goods Sold ($);

# Total Expenses: Total Expenses to get the Product to Sell ($);

# Marketing: Expenses in Marketing ($);

# Inventory: Inventory Value of the Product in the Sale Moment ($);

# Budget Profit: Expected Profit ($);

# Budget COGS: Expected COGS ($);

# Budget Margin: Expected Profit + Expected Total Expenses ($) OR Expected Sales - Expected COGS ($);

# Budget Sales: Expected Value Acquired in Sales ($);

# ProductID: Product ID;

# Date: Sale Date;

# Product Type: Product Category;

# Product: Product Description;

# Type: Type;

#####1.2. DATOS CUANTITATIVOS

####

# Area code: Entero

# Profit: Entero

# Margin: Entero

# Sales: Entero

# COGS: Entero

# Total Expenses: Entero

# Marketing: Entero

# Inventory: Entero

# Budget Profit: Entero

# Budget COGS: Entero

# Budget Margin: Entero

# Budget Sales: Entero

# ProductId: Entero

###

#####1.3. DATOS CUALITATIVOS

####

#State: Cadena

#Market: Cadena

#Market Size: Cadena

#Product Type: Cadena

#Product: Cadena

####

# 2. Preprocesamiento de Datos: Descripción detallada de los pasos tomados para limpiar y preparar los datos para el análisis utilizando Tidyverse.

datos <- read\_delim("sales.csv", delim = ",") # Usar read\_delim con el delimitador especificado

# Inspeccionamos las primeras filas de los datos

head(datos)

# Resumen de estructura de datos (imprime algunos datos y tipos de datos)

glimpse(datos)

# Inspeccionar y limpiar los datos si es necesario (imprime la estructura y los tipos de datos)

str(datos)

# 3. Análisis Exploratorio de Datos: Implementación de códigos en R para explorar los datos. Esto puede incluir visualizaciones de datos, medidas de tendencia central, etc. importación de datos

# 3.1. resumen general/total estadísticos de todos los datos presentes en el dataset

summary(datos)

# 3.2. Comprobar si hay valores faltantes en las columnas importantes

na\_count <- sapply(datos, function(x) sum(is.na(x)))

na\_count

# 3.3. Comprobar si hay valores faltantes en las columnas importantes seleccionadas que exige el informe.

sum(is.na(datos))

sum(is.na(datos$Profit))

sum(is.na(datos$Sales))

sum(is.na(datos$COGS))

# 3.4. Histograma de Profit

ggplot(datos, aes(x = Profit)) +

geom\_histogram(bins = 30, fill = "blue") +

ggtitle("Histograma de Profit")

# 3.5. Boxplot de Sales

ggplot(datos, aes(y = Sales)) +

geom\_boxplot(fill = "green") +

ggtitle("Boxplot de Sales")

# 3.6. Boxplot de Total Expenses por Market

ggplot(datos, aes(x = Market, y = `Total Expenses`)) +

geom\_boxplot(fill = "cyan", color = "black") +

ggtitle("Boxplot de Total Expenses por Market") +

xlab("Market") +

ylab("Total Expenses")

# 4 Preprocesamiento de Datos (formateo de solamente fecha, ya que la fecha incluye horas, minutos y segundos)

datos <- datos %>%

mutate(fecha = as.Date(fecha, format = "%d-%m-%Y")) %>% # Convertir la columna 'fecha' al formato de fecha (no incluye hh:mm:ss)

filter(!is.na(precio)) %>% # Eliminar filas donde el precio es NA

arrange(desc(precio)) # Ordenar los datos por precio de forma descendente